Building Informal Inference in Grade 7

Jane Watson
University of Tasmania
<Jane.Watson@utas.edu.au>

Julie Donne
University of Tasmania
<Julie.Donne@utas.edu.au>

This study reports on the second phase of a design experiment involving classroom implementation of a sequence of four lessons introducing informal inference supported by *TinkerPlots* software to a grade 7 class. A Beginning Inference Framework was used as an implicit foundation for the teachers and as an explicit rubric for assessing students' observed outcomes. Outcomes were judged in relation to saved *TinkerPlots* files annotated with student-completed text boxes and to individual interviews with 12 of the students.

In 2006, as part of a larger professional learning research project, Jenny (pseudonym), a grade 7 teacher in a rural district school (K-10), undertook a case study related to introducing her class to *TinkerPlots* graphing software for middle schools (Konold & Miller, 2005). The case study evolved into a design experiment adapting lessons to cover elements of a Beginning Inference Framework, derived from a model suggested by Pfannkuch (2006). Data collected in the form of *TinkerPlots* files from four sessions were analysed in relation to the Beginning Inference Framework to document students' observed progress in taking up the elements of informal inference (Watson, 2007). The key aspects of the initial intervention included the evaluation of the extent to which the elements of the framework were observed in student output.

The 2006 case study and the subsequent 2007 case study described in this report arose from the desire of the statistics education research community to provide a meaningful bridge to formal inference, which many students will meet at the senior secondary or tertiary level. As well there is the desire to provide students who do not go on to formal statistics with intuitions about the inferential process without the theoretical assumptions and more complex mathematics required in formal statistics courses. The school curriculum provides direction on some of the ingredients required, such as data representation in graphs and data reduction with averages, but often does not signal the purpose of decision making with uncertainty based on samples representing populations. The National Council of Teachers of Mathematics (2000) includes "develop and evaluate inferences and predictions" in its *Standards* at all levels but there is concern on the part of statistics educators about how this is implemented, especially in acknowledging the uncertainty in the evaluation process.

As part of her wider work with senior secondary teachers in New Zealand, Pfannkuch (2006) set up a framework involving eight elements for developing informal inference based on box plots, which were a significant representational form in the New Zealand curriculum. Two other inputs influenced the adaptation of Pfannkuch's model for the study described here. First was the work of Bakker, Biehler, and Konold (2005), which concluded that box plots placed demands on students in terms of proportional reasoning that were beyond the understanding of most middle school students. This was especially true since most representations of box plots appeared without the inclusion of the data that they were summarising. Second was the development of the *TinkerPlots* software and its provision of a tool called the hat plot, which is a simplified version of a box plot. The hat plot (the default form) appears "above" the data (if plotted horizontally) with its crown situated over the middle 50% of the data and its brims over the lowest and highest 25% of the data. The median does not appear in the hat and hence the data are likely to be considered in "thirds," these being the middle cluster and two extremes. The availability of *TinkerPlots* as part of the project hence led to the adaptation of the Pfannkuch framework to the one in Table 1.

Table 1

Beginning Inference Framework (adapted from Pfannkuch, 2006)

Element	Description
Hypothesis Generation	Reasons about trends (e.g., differences)
Summary	Summarizes the data using the graphs and averages produced in *TinkerPlots*
Shift	Compares one hat plot with the other/s referring to change (shift)
Signal/Centre	Refers to (and compares) information from the middle 50% of the data
Spread	Refers to (and compares) spread/densities locally and globally
Sampling	Considers sampling issues such as size and relation to population
Explanatory/Context	Understands context, whether findings make sense, and alternative explanations
Individual Case/s	Considers possible outliers and other interesting individual cases

The results of the 2006 case study (Watson, 2007) were encouraging in that at each of the four data collections, more of the elements of the framework were employed by students, and after a 3½-month break students were able to engage with the elements presented in a structured format including graphs prepared in *TinkerPlots*. Concerning to the research team, however, was the difficulty students had in linking the ideas of sample and population. It was felt that this difficulty may have been related to discussions held with students when their data were collected and questions asked about them and the middle school students in their school. It was not until later that the larger population of "all" middle school students in the state or nation was introduced. The students appeared to have difficulty appreciating why these larger questions were of interest or important. This was one of the main features of instruction that was intended to be amended in the current case study. Within the context described, the research questions for this study are hence the following: What are the observed learning outcomes for grade 7 students in relation to beginning inference in a learning environment supported by a Beginning Inference Framework, a software package for data handling, and revised implementation strategies? How do the learning outcomes suggest further changes to the framework, the implementation, or the interaction with the software?

Methodology

This study is seen as the second phase of a design experiment (e.g., Cobb, Confrey, deSessa, Lehrer, & Schauble, 2003; The Design-Based Research Collective, 2003). The characteristics include the evolving theoretical framework for beginning inference, the interactive nature of the intervention (Jenny, a teacher-researcher (T-R), and the first author), the variety of data sources employed, and the adaptation of the intervention potentially to make suggestions for future research.

Participants. The case study was based in a grade 7 class (12-13 years old) of 15 students (4 other students opted not to be involved with the study and were transferred to other classes for the time of the lessons described here). Jenny, the T-R, and the first author had been involved in the earlier case study (Watson, 2007), which had included professional learning for Jenny and a close classroom collaboration of Jenny and the T-R.

Procedure and data collection. After initial planning, the students had been given time to explore *TinkerPlots*. *Lesson 1* introduced an investigation evolving around the hypothesis of an 81-year-old man that in the population at large males have faster reaction times than females. Students collected data on their right and left hand reaction times as a sample using the Australian Bureau of Statistics *CensusAtSchool* web site. Students entered data in *TinkerPlots* and created graphs to explore the hypothesis. Comments were entered in text boxes. In *Lesson 2*, the T-R led a discussion with Jenny at the computer using a *TinkerPlots* file on homework data. This covered the various tools available in *TinkerPlots* and the students then used these tools to explore their class's data set from the previous session. In *Lesson 3*, students were introduced to random samples of 20 or 200 grade 7 students reaction times collected from the *CensusAtSchool* web site. In *Lesson 4* students had access to a random sample of size 200 grade 5 and 12 students from the *CensusAtSchool* web site. All lessons lasted between 1.5 and 2 hours and videotapes were made of Lessons 2, 3, and 4. These provided audio but not always video records of events. *TinkerPlots* outputs were collected from all students present at each lesson.

The subsequent student interviews several weeks later introduced the students to new data sets already entered into *TinkerPlots* data cards. Students were asked to answer and discuss questions with the interviewer (one of the authors) rather than to write responses in text boxes. Three protocols were used in the interviews. The Comparing Groups protocol (Watson & Moritz, 1999) asked students to compare four pairs of classes on the basis of their spelling scores and decide which class had done better. The first three pairs of classes were of small equal sizes, whereas the fourth pair was not of equal size. The second data set consisted of 16 data cards with the names, ages, weights, eye colours, favourite activities, and numbers of fast food meals eaten per week of 16 students aged between 8 and 18 (Chick & Watson, 2001). Students were asked to explore the cards, suggest interesting hypotheses, and provide plots with evidence to support or refute the hypotheses. The third protocol was based on a *TinkerPlots* data set containing the heights of 136 children at age 2, 9, and 18 (Watson, 2007). Students were asked to form hypotheses about the difference in heights for boys and girls at the three times based on stacked dot plots provided for each year, separated by gender and including hat plots (the scales were different on each of the three plots).

Analysis. Following the method of analysis of the previous case study, the Beginning Inference Framework was the basis for analysis of the three taped class sessions, the student *TinkerPlots* output from each of the four sessions, and the individual interviews. For each lesson and the interviews, matrices were created to document students' work (Framework elements x students). Judgments of outcomes were based on the number of elements employed and the degree to which they were related to each other. The relationships were categorised based on the SOLO Taxonomy (Biggs & Collis, 1982; Pegg, 2002) as employed by Watson (2007). Prestructural responses, reflecting no elements of the Framework, were not observed in this study due to the scaffolding of the sessions and the collection of *TinkerPlots* output from students. Outcomes were judged to be Unistructural (U) if isolated comments or plots were saved, based solely on classroom discussion. Responses that added extra elements of the Framework in a serial fashion to the comments and annotations to plots were judged to be Multistructural (M). Relational (R) responses were those that combined several elements in the text comments to reach integrated conclusions in the *TinkerPlots* output. For the interviews judgments were made based on the overall use made of the elements across the three protocols, suggesting the degree to which the students had internalised the experiences that had taken place across the four teaching sessions, with similar criteria to those above being employed. All quotes have been corrected to fix spelling and grammatical errors.

Results

The results are presented in three parts. First the elements observed in the lessons are documented as evidence for the experiences of the students in the four lessons. Next, the observed outcomes from the students' *TinkerPlots* files are summarised for the four lessons. Finally, the observed outcomes for the individual interviews are presented and a summary given for the students over the five data collections.

Lesson summaries. Table 2 contains annotations in relation to each of the eight elements of the Beginning Inference Framework for the three videotaped sessions. Evidence for Lesson 1 is noted with Lesson 2 and was gleaned from discussions with the T-R and student output. The only elements not addressed specifically in a session were Hypothesis Generation and Explanatory/Context in Lesson 2, which were strongly addressed in Lesson 1, and Explanatory/Context in Lesson 4, as it had not changed substantially over the four sessions.

Table 2

Beginning Inference Framework Elements Addressed Across Lessons

Element	Lessons 1 and 2 – Introduction, Class Data	Lesson 3 – ABS grade 7 data (20, 200)	Lesson 4 – ABS grade 5 and 12 data (200)
Hypothesis Generation	Jim, 81 years old, hypothesis boys have faster reaction times than girls [focus Lesson 1]	Review and reminder of alternatives; general discussion of hypotheses	Suggestions: grade 5 faster than grade 12; some boys faster than some girls
Summary	Ranges, middles, n, percent	S: "dots all over the place"; boys faster left but right about the same	Clusters, stacking, concentrations, reference lines
Shift	More people on left; ranges the same; ranges different	Not much shift in the data	Endpoints of crowns of hats notes; "girls start later"
Signal/Centre	Mean, median, hat: 50%, 25%, 25%	Reminder hat is middle 50%; mean, median	Hat middle 50%; median; "girls end of 50% for boys"
Spread	Range, range of middle 50%; fixing scale to compare spread	Scale, sensible ranges; clumped; spread out; range of middle	Range/scale; girls wider crown, more spread; top & bottom 25%, clusters, pencil
Sampling	What about grade 8, equal numbers, all girls/boys in class?	Samples of 20, 200; equal number of boys/girls?	Missing data; small/large data sets and outliers
Explanatory/ Context	Reasons for possible difference [focus Lesson 1]	Continued as earlier	[Little extra] mainly hypotheses and evidence
Individual Case/s	Outlier [one class member], one left-handed student	Review of outliers; "flukes", 0 as outlier	Outliers, flukes, specific values in data set

Student outcomes – Lesson 1. Of the 15 students in Lesson 1, 10 produced *TinkerPlots* files with at least the scaled data set in stacked format. One produced a plot separated by gender showing 10 males and 9 females. Three produced 2-way plots with four bins showing gender and right hand reaction time in two groups, 0.24 – 0.35 s and 0.36 – 0.48 s. One produced a plot of gender by eight subgroups of right hand reaction time. S4's response was considered typical of Multistructural responses and included five elements in comments with a stacked dot plot: Hypothesis Generation (who is faster), Summary (right hand plot, boys faster by one microsecond), Sampling ("we only did grade 7"), Explanatory/Context (most right-handed, one left-handed), and Individual Cases (two fastest, third fastest named). Unistructural responses were similar to S2, who stated the hypothesis, noted that data had been collected and graphs made, but no conclusion was drawn. Of the responses, 8 were judged to be Unistructural, and 7, Multistructural.

Student outcomes – Lesson 2. Only 10 students were present at the second session. Students analysed their class data with more of the tools available in *TinkerPlots* and saved as many as four new plots. Only newly added text was considered in deciding the level of response. Only one student did not produce a plot separating boys and girls; eight looked at both left and right hand times by gender. Some accounts were mainly descriptive of the procedures followed in creating the plots using the tools. Eight removed the outlier from the left hand times; one kept it "because I thought that it was important to leave her in so it's exact" [S15]. Most of the students used reference lines to detail values and six used hat plots in at least one of their graphs. Comments summarising the graphs displayed a range of uncertainty of language: for example, S13 declared "the boys are faster than the girls," whereas S4 concluded, using reference lines and medians, "some boys are faster than girls sometimes." Three students included an aspect of sampling of "our class" and four described the spread of their plots, for example with ranges of the crowns of the hats [S12]. S1 was however confused that a wider crown meant more data rather than greater spread for a fixed percent of the data. In providing responses such as these, eight of the students were judged to provide Multistructural responses. The other two responses were considered Unistructural because in one case nothing of substance was added to the previous week's work and in the other the student recorded contradictions that precluded understanding the conclusions drawn.

Student outcomes – Lesson 3. In this session 13 students were present. With the choice of considering a randomly selected data set from the ABS *CensusAtSchool* site with either 20 or 200 grade 7 students, 11 chose the set of 200, although later many suggested that is was difficult for them to work with such a large data set. Three students did not delete outliers, despite much class discussion; two deleted high values but not zeroes. Three students did not state new hypotheses or questions to explore. All students considered gender by reaction time, 11 for both hands and two for one hand only. Eleven students produced hat plots but did not mention change in their location for the two sexes. Seven students discussed some other aspect of the middle 50% of the data, either using reference lines or explicitly noting the "middle half" or the "50% majority," whereas four considered spread by giving values for the range of the crowns. Overall the graphs were summarised to the extent of saying "they showed" support for the hypothesis. Generally the students struggled to document evidence in their text boxes to support their conclusions. Although most of the students learned to handle the larger data set with more outliers, in many cases the subtleties of the positions of the hats made decisions difficult. Nine of the students were judged to present Unistructural responses, five of whom had been absent the previous week; of the other four comments were contradictory or did not describe evidence in support of conclusions. Figure 1 shows the two plots produced by S7 who gave values for the ranges of the crowns of the hats and concluded "in the right hand times the girls and boys are about the same in times … but in the left hand times the boys were faster." This was typical of the four Multistructural responses.

Student outcomes – Lesson 4. Of the 13 students present, four had missed Lesson 2 and two had missed Lesson 3. In this session students were aided by a review of the previous session's analysis and most appeared to contribute to the discussion. When given a new randomly selected data set of 200 students in grades 5 and 12, they immediately considered gender with reaction time, without consideration of grade. When reminded of this, students then looked at grade without consideration of gender. They did not have the experience with *TinkerPlots* or time to consider how both attributes could be considered together. The students, however, consolidated the procedures using the *TinkerPlot* tools. All students except one placed hat plots on some or all of their graphs and the one student used reference lines instead.

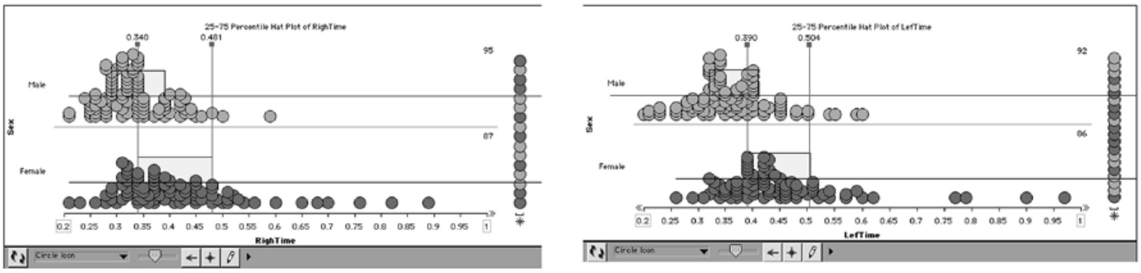

Figure 1. Random sample of grade 7 data for left and right hand reaction time by gender.

The assessment of responses was based on how far students could progress given the scaffolding provided. Three students wrote little text to explain their graphs and did not summarise the information to support a hypothesis. One student based the discussion only on individual values at the extremes of the plots. The responses of these four students were considered to be Unistructural in the context; three of these students had missed one earlier session. Eight students produced responses judged to be Multistructural in that they made reasonable hypotheses about gender and grade separately with respect to reaction time and supported these with reference to their plots. None of the students addressed aspects of sampling, context, or specific shifting of hats, except for S8 who noted for one plot, "boys also have the middle 50% nearer the beginning but the girls are further back." S10 was considered to produce a Relational response in that he made comments reflecting all eight elements of the Beginning Inference Framework, for example, for Shift, "the middle 50% range in the males starts at a lower time than females in the right hand, the end of the 50% range in males and females [is] at the same place"; for Spread, "the slope of the males in the right had has a more jagged slope and the females have a smooth slope"; for Sampling, "year 5 had more people … also more outliers or incomplete data in grade 5." S10 also noted the mixed gender within grades.

Student interviews. The specific elements of the Beginning Inference Framework were not brought to the attention of the 12 students who were interviewed, except that they were asked to generate hypotheses in the second and third protocols. Of interest was the degree to which the elements were integrated into the comments made by the students to the interviewers. Four students showed little awareness of the task of setting hypotheses; three of these had missed one session. The other eight were successful in one or more contexts. Two, including one of the previous four struggled with summarising plots to reach conclusions. Seven students discussed shift, without using the term, for hat plots. There was some confusion on the signal in the central 50% of the data but six could make reasonable comments. All students mentioned spread and looked at individuals (or individual bins with one or two entries) in one or more of the protocols. Seven responses considered Explanatory/Context, for example giving advice about eating fast foods or discussing cases of growth for the final protocol. Only two students discussed the need for larger samples throughout the interview. In considering the overall adoption of the elements of the Beginning Inference Framework, it was judged that four students' responses were Unistructural in focusing overwhelmingly on individual aspects of the data sets rather than aggregate properties represented or representable in plots. Six responses were considered Multistructural in displaying many of the elements at various points of the interview, whereas two were considered Relational in integrating the elements to create meaningful arguments. Table 3 contains the number of elements of the Framework discussed and how they were structured according to SOLO levels for each student for each data collection. The average number of elements included in the *TinkerPlots* output increased from 3.2 in Lesson 1, to 3.9 in Lesson 2, and to 4.8 in Lesson 3, before dropping slightly to 4.6 in Lesson 4. The average number of elements observed in the transcripts of student discussion rose to 5.7 in the interviews.

Table 3

Number of Elements of Framework and SOLO level for Students at Five Times

	Lesson/Interview									
Student	L1 No. Elem.	L1 SOLO	L2 No. Elem.	L2 SOLO	L3 No. Elem.	L3 SOLO	L4 No. Elem.	L4 SOLO	Int No. Elem.	Int SOLO
S1	2	U	5	M	–	–	5	M	6	M
S2	2	U	–	–	4	U	4*	M	3+	U
S3	2	M	–	–	4	U	–	–	3+	U
S4	5	M	3	M	2	U	4	M	6+	M
S5	4	U	–	–	3	U	4	U	–	–
S6	4	U	4	M	–	–	4	U	6+	U
S7	4	U	6	M	7	M	5	M	8	M
S8	3	M	3	U	6	M	5	M	5+	M
S9	4	M	4	U	5	U	4*	M	7	M
S10	3	M	3	M	7	M	8	R	7	R
S11	2	U	–	–	5	U	4	M	–	–
S12	2	U	3	M	4	U	3	U	5	U
S13	3	U	4	M	6	U	–	–	–	–
S14	5	M	–	–	4	U	5	U	6	M
S15	3	M	4	M	5	M	5	M	6+	R

*Plus one element that was used inappropriately + Weak use of another element

Discussion

In initiating a second series of lessons in a design experiment framework, one aim was to observe the effect of transposing the introduction of populations and samples from after the collection of student class data to the very beginning of the lesson sequence. This was accomplished by introducing a population-based hypothesis with class discussion of how the sample data from the class might assist in supporting the hypothesis or refuting it. Comments were made about hypotheses not being "right or wrong" but questions to be investigated by collecting evidence. Although appearing to appreciate and participate in the class discussion (for example, observed in audio extracts on the lesson videos), little explicit evidence of this was volunteered in the text boxes or in the interviews. This is probably related to the specific interest in the *TinkerPlots* features and the relative ease with which they could be described. The authors did not ask specific questions about population and sampling in the interviews because of the desire to find out what students would contribute on their own initiative. The second protocol with 16 data cards provided an opportunity for students to make comments on the need for a larger sample or what might happen for the population at large.

It is likely that the drop in SOLO levels, despite the increase in average number of Framework elements observed in Lesson 3, is related to the introduction of a new, and for most students much larger, data set, as well as to the fact that some students had been absent for Lesson 2. The slight drop in number of Framework elements observed in Lesson 4 may be related to the extra cognitive load of considering both gender and grade level. The improved SOLO levels may relate to some students beginning to put together the ideas of combining evidence to reach a conclusion. Increasing numbers of elements employed in the interview may have resulted from the many opportunities provided to students but the continued presence of Unistructural responses suggests that some students still struggled with more than considering single aspects that resulted from employing *TinkerPlots* tools.

Overall the authors conclude that in terms of grade 7 students assimilating concepts of populations and sampling along with the other elements of the Beginning Inference Framework, little is gained by introducing the "big picture" of populations first rather than later in an investigation sequence. The lack of spontaneous intuitive consideration of populations may be associated with students' continued focus at this age on themselves and their immediate environment or it may take much longer with more experiences than were possible in this case study to build appropriate intuitions about sampling.

Although the Beginning Inference Framework was adopted in a context of replacing box plots with hat plots, not all students continued to use hat plots in Lesson 4 or the interview, some preferring data in bins and others using arbitrary reference lines. Whether this was again a function of lack of experience in seeing the usefulness of hats in various contexts is unknown. In Pfannkuch's (2006) study with older students box plots were the only representation provided for interpretations (without actual data values), whereas in this case study hats were one of a number of tools available to summarise a plot of data values. For beginners it seems reasonable to provide a range of tools, such as available with *TinkerPlots*, with the intention of allowing students to build intuitions that will assist in the transition into more formal inference methods and the use of box plots in later years. Further interventions over a number of years will be needed to test these ideas.

Acknowledgment. The study was funded by ARC projects LP0560543 and LP0669106.

References

Bakker, A., Biehler, R., & Konold, C. (2005). Should young students learn about box plots? In G. Burrill & M. Camden (Eds.), *Curricular development in Statistics education: International Association for Statistical Education (IASE) Roundtable, Lund, Sweden, 28 June-3 July 2004* (pp. 163-173). Voorburg, The Netherlands: International Statistical Institute.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Chick, H. L., & Watson, J. M. (2001). Data representation and interpretation by primary school students working in groups. *Mathematics Education Research Journal*, *13*, 91-111.

Cobb, P., Confrey, J., deSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13.

Konold, C., & Miller, C. D. (2005). *TinkerPlots: Dynamic data exploration*. [Computer software] Emeryville, CA: Key Curriculum Press.

Pegg, J. E. (2002). Assessment in mathematics: A developmental approach. In J. M. Royer (Ed.), *Mathematical cognition* (pp. 227-259). Greenwich, CT: Information Age Publishing.

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal*, *5*(2), 27-45.

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

The Design-Based Research Collective (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5-8.

Watson, J. M. (2007, August). *Facilitating beginning inference with TinkerPlots for novice grade 7 students*. Refereed paper presented at the Fifth International Forum for Research on Statistical Reasoning, Thinking and Literacy, University of Warwick, Coventry, UK.

Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, *37*, 145-168.